Weakly informative priors

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

23 Apr 2014

Collaborators (in order of appearance):
Gary King, Frederic Bois, Aleks Jakulin, Vince Dorie, Sophia Rabe-Hesketh, Jingchen Liu, Yeojin Chung, Matt Schofield, Ben Goodrich, . . .

Weakly informative priors for . . .

- Two applied examples
  1. Identifying a three-component mixture (1990)
  2. Population variation in toxicology (1996)
- Some default priors:
  3. Logistic regression (2008)
  4. Hierarchical models (2006, 2011)
  5. Covariance matrices (2011)
  6. Mixture models (2011)

1. Identifying a three-component mixture

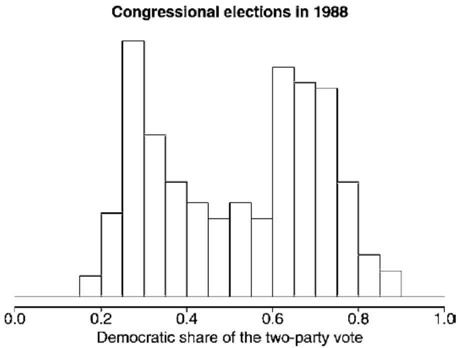


Figure 1. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988. Only districts that were contested by both major parties are shown here.

- Maximum likelihood estimate blows up
- Bayes posterior with flat prior blows up too!

Priors!

- Mixture component 1: mean has $N(-0.1, 0.1^2)$ prior, standard deviation has inverse-$\chi^2_4(0.1^2)$ prior
- Mixture component 2: mean has $N(+0.1, 0.1^2)$ prior, standard deviation has inverse-$\chi^2_4(0.1^2)$ prior
- Mixture component 3: mean has $N(0, 0.3^2)$ prior, standard deviation has inverse-$\chi^2_4(0.2^2)$ prior
- Three mixture parameters have a Dirichlet $(19, 19, 4)$ prior

Separating into Republicans, Democrats, and open seats

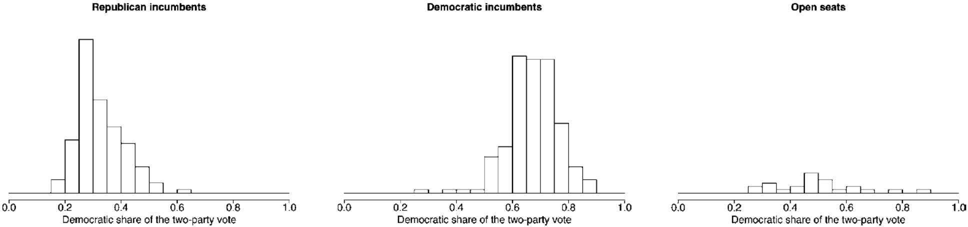


Figure 2. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988, in Districts With (a) Republican Incumbents, (b) Democratic Incumbents, and (c) Open Seats. Combined, the three distributions yield the bimodal distribution in Figure 1.

- Beyond "weakly informative" by using incumbency information

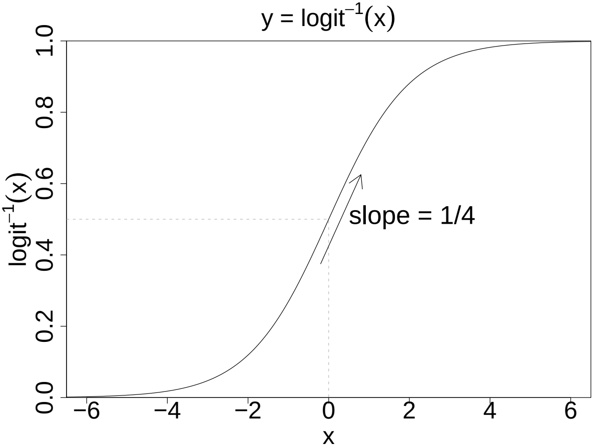
2. Weakly informative priors for population variation in toxicology

- Pharamcokinetic parameters such as the "Michaelis-Menten coefficient"
- Wide uncertainty: prior guess for θ is 15 with a factor of 100 of uncertainty, $\log\theta \sim \mathrm{N}(\log(16), \log(10)^2)$
- Population model: data on several people j, $\log\theta_j \sim \mathrm{N}(\log(16), \log(10)^2)$????
- Hierarchical prior distribution:
  - $\log\theta_j \sim \mathrm{N}(\mu, \sigma^2), \quad \sigma \approx \log(2)$
  - $\mu \sim \mathrm{N}(\log(15), \log(10)^2)$
- Weakly informative

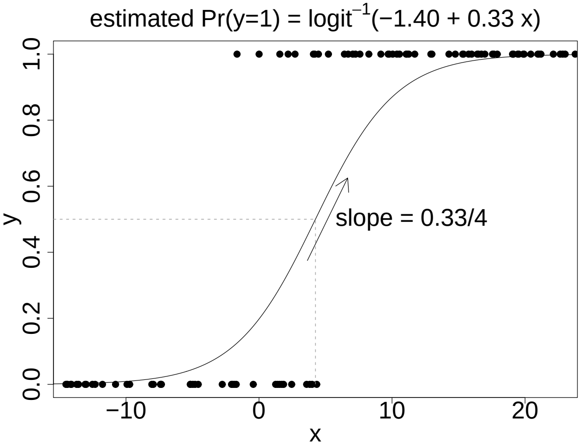
Concepts

- Wip instead of noninformative prior or informative prior
- You're using wips already!
- Prior as a placeholder
- Model as a placeholder
- "Life is what happens to you while you're busy making other plans"
- Full Bayes vs. Bayesian point estimates
- Hierarchical modeling as a unifying framework

3. Logistic regression



A clean example



The problem of separation

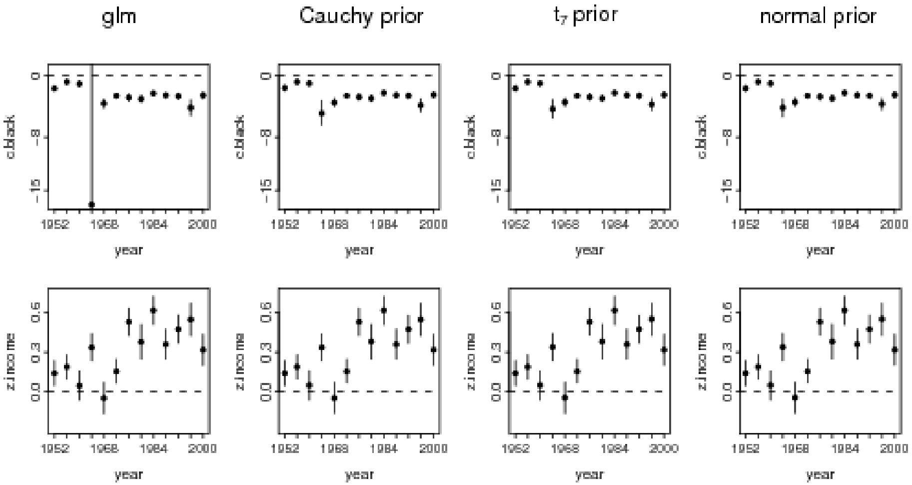
Separation is no joke!

```
glm (vote ~ female + black + income, family=binomial(link="logit"))

1960                                    1968
            coef.est coef.se                        coef.est coef.se
(Intercept) -0.14     0.23              (Intercept)  0.47     0.24
female       0.24     0.14              female      -0.01     0.15
black       -1.03     0.36              black       -3.64     0.59
income       0.03     0.06              income      -0.03     0.07

1964                                    1972
            coef.est coef.se                        coef.est coef.se
(Intercept)  -1.15     0.22             (Intercept)  0.67     0.18
female       -0.09     0.14             female      -0.25     0.12
black       -16.83   420.40             black       -2.63     0.27
income        0.19     0.06             income       0.09     0.05
```

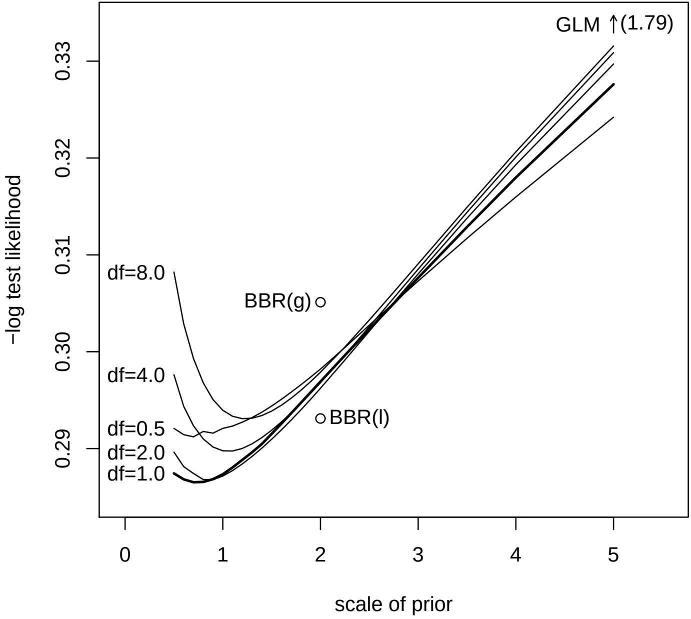
Regularization in action!



Weakly informative priors for logistic regression

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Expected predictive loss, avg over a corpus of datasets



What does this mean for YOU?

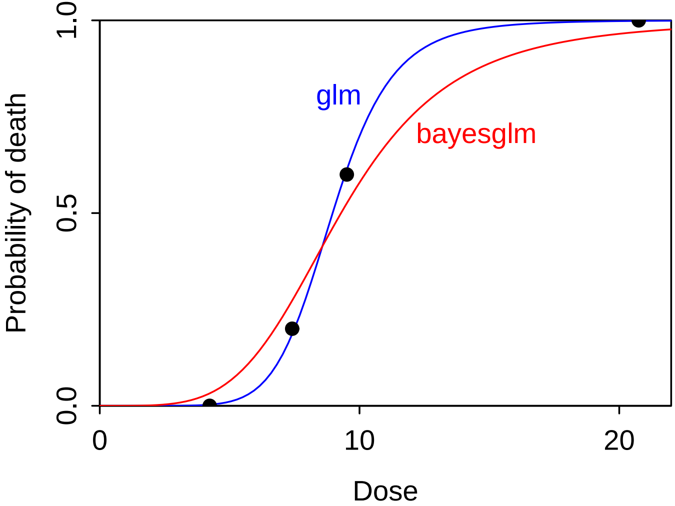
- Sparse data
- Big Data need Big Model

Another example

Dose	#deaths/#animals
-0.86	$0/5$
-0.30	$1/5$
-0.05	$3/5$
0.73	$5/5$

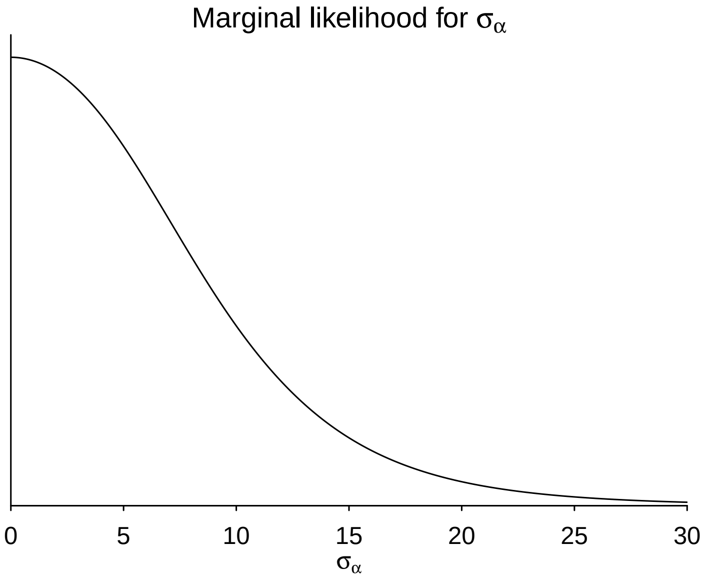
- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9

Maximum likelihood and Bayesian estimates



- Which is truly conservative?
- The sociology of shrinkage

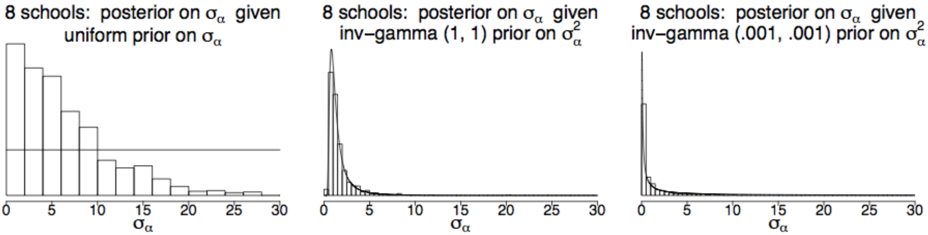
4. Inference for hierarchical variance parameters



Hierarchical variance parameters: 1. Full Bayes

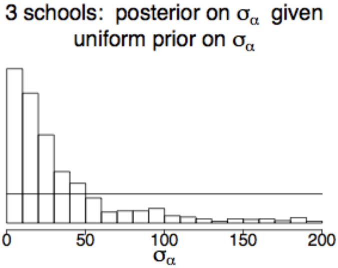
- What is a good "weakly informative prior"?
  - $\log \sigma_\alpha \sim \text{Uniform}(-\infty, \infty)$
  - $\sigma_\alpha \sim \text{Uniform}(0, \infty)$
  - $\sigma_\alpha \sim \text{Inverse-gamma}(0.001, 0.001)$
  - $\sigma_\alpha \sim \text{Cauchy}^+(0, A)$
- Polson and Scott (2011):
  "The half-Cauchy occupies a sensible 'middle ground' . . . it performs very well near the origin, but does not lead to drastic compromises in other parts of the parameter space."

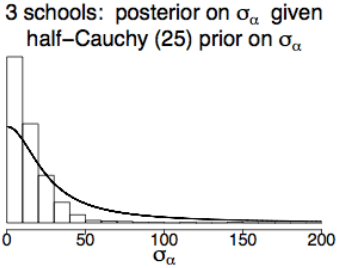
Problems with inverse-gamma prior



- Inv-gamma prior cuts off at 0

Problems with uniform prior



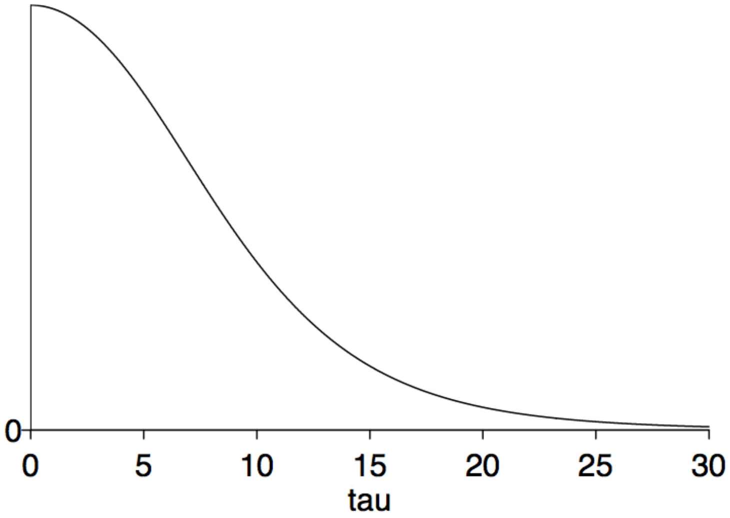


- Uniform prior doesn't cut off the long tail

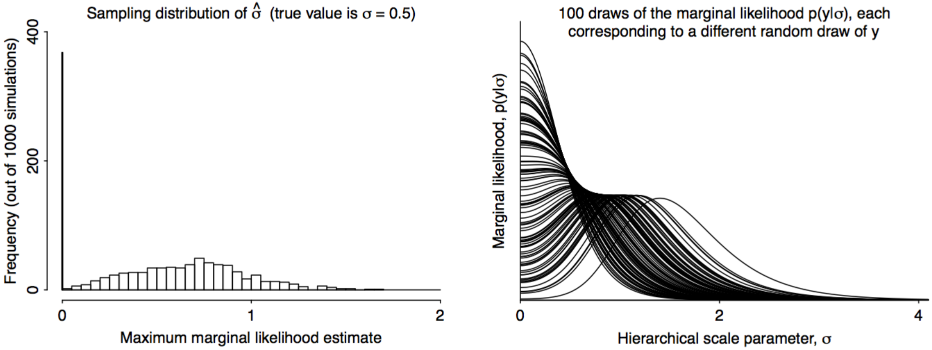
Hierarchical variance parameters: 2. Point estimation

- [Estimate $\pm$ standard error] as approximation to full Bayes
- Point estimation as goal in itself
- Problems with boundary estimate, $\hat{\sigma}_\alpha = 0$

The problem of boundary estimates: 8-schools example



The problem of boundary estimates: simulation



The problem of boundary estimates: simulation

- Box and Tiao (1973):

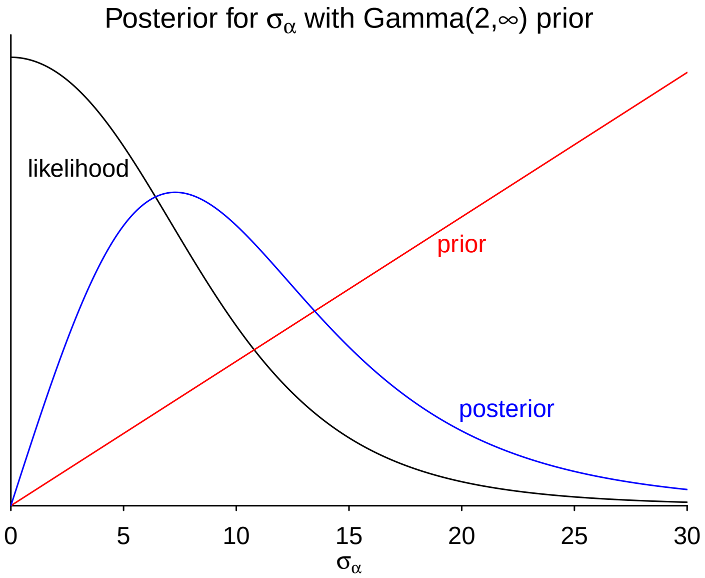
> The second data set we consider illustrates the case where the between-batches mean square is less than the within-batches mean square. These data had to be constructed for although examples of this sort undoubtedly occur in practice they seem to be rarely published. The model in (5.1.3) was used to generate six groups

- All variance parameters want to become lost in the noise
- When does it hurt to estimate $\hat{\sigma}_\alpha = 0$?

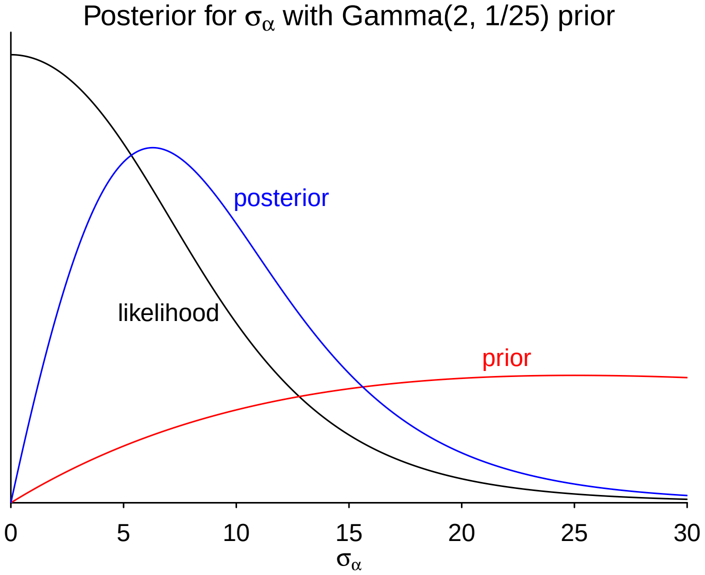
Point estimate of a hierarchical variance parameter

- Desirable properties:
  - Point estimate should never be 0
  - But ... no nonzero lower bound
  - Estimate should respect the likelihood
  - Bias and variance should be as good as mle
  - Should be easy to compute
  - Should be Bayesian

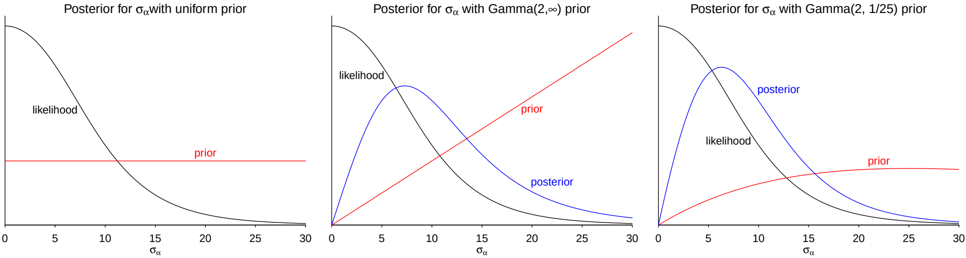
Boundary-avoiding point estimate!



Boundary-avoiding weakly-informative point estimate



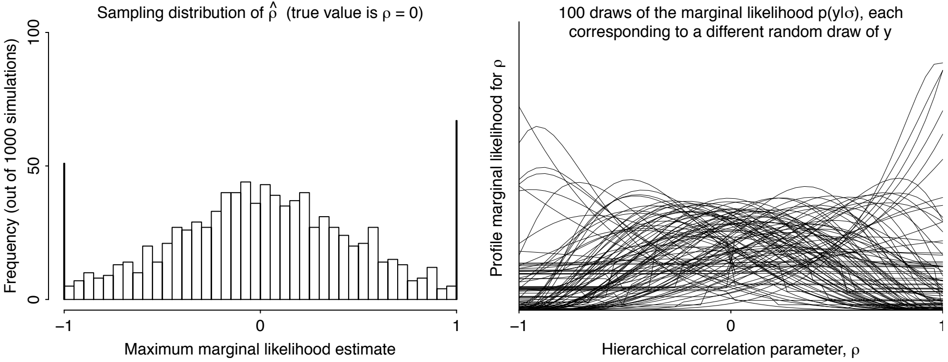
Gamma (not inverse-gamma) prior on σ_α



Posterior mode is shifted at most one standard error from the boundary

5. Boundary estimate of group-level correlation

Point estimates of correlation $\hat{\rho}$ from a hierarchical varying-intercept, varying-slope model:



- Boundary-avoiding prior: $\rho \sim \text{Beta}(2, 2)$
- Better statistical properties than mle

Weakly informative priors for covariance matrix

- Boundary-avoiding prior
  - Wishart (not inverse-Wishart) prior
  - Generalization of gamma
- Full Bayes:
  - Scaled prior on Diag*Q*Diag
- Connection to prior on regression coefficients

6. Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don't "look" like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative prior: use a hierarchical model for the scale parameters

General theory for wips

- (Unknown) true prior, $p_{\mathrm{true}}(\theta) = N(\theta|\mu_0, \sigma_0^2)$
- Your subjective prior, $p_{\mathrm{subj}} = N(\theta|\mu_1, \sigma_1^2)$
- Weakly-informative prior, $p_{\mathrm{wip}} = N(\theta|\mu_1, (\kappa\sigma_1)^2)$, with $\kappa > 1$
- Tradeoffs if κ is too low or too high

Specifying wips using nested models

- Thiago Martins, Daniel Simpson, Andrea Riebler, Havard Rue, Sigrunn Sorbye
- Large model has parameter θ
- Define $p(\theta)$ with reference to some `base model` such as $\theta = 0$
- Penalized complexity priors
- User-defined scaling

What have we learned?

- Models need structure but not too much structure
- Conservatism in statistics
- Priors for full Bayes vs. priors for point estimation
- Formalize the losses in supplying too much or too little prior info